Are those California's Houses in 1990 located near the Ocean more Expensive, on Average?

Luis Rojas S.

December 18, 2020

1 Introduction

This paper analyzes empirically data from California in order to test the hypothesis that those houses locates near the coast tend to be more expensive, on average. We suggest that those houses near the ocean tend to be more expensive, on average, but this may be the case that these properties are bigger or located near important economic centers (like downtown Los Angeles) and other factors that may move the price along the location.

To purge the house price from these factors we add a set of control variables and later with the help of maps visualize where the most expensive houses are concentrated. This tool helps to have a quick look of the distribution of houses in California, moreover we can appreciate an important finding: some of the houses that are located near the ocean are in fact cheap. We then suggest that these properties tend to be cheaper due to its location near to important geological failures. One of the most dangerous of them is called the San Andreas fault zone, so we argue that a property investment near a zone where earthquakes are common drives the price down since there is an added risk when purchasing the house. Another map using the geological failures (earthquakes) known up to 1990 may help to clarify this argument. Previous research have added valuable information to this field providing with valuables datasets and models that help to understand better the determinants of the price of a house.

Zietz J., Emily Norman Zietz E., and Sirmans G. (2007) mention that house prices and the characteristics of a house share a close relationship. They also suggested that some characteristics of

a house tend to be priced differently depending on the economic quantile of the potential buyer. The authors use OLS regression with quantile regressions and other types of econometric models (such as 2SLS), and also mention controls for spatial autocorrelation. This paper uses an extensive number of controls that may help to get better results for estimation purposes, differently we are using a significant smaller number of variables in order to test our hypothesis.

This paper narrows the current research in this field by considering the location to the ocean as a potentially predictor of the price of a house. To test the hypothesis that those houses located near the ocean tend to be more expensive on average we use the California 1990 census that provides observations at the district level for the 58 counties in California; this dataset includes observations for the demographic level and characteristics of the houses at this district.

2 The California 1990 Census Data.

The California 1990 census collects 20,640 observations among 10 variables seized at the district level. The categorical variable *ocean_proximity* records the position of the district to the ocean; in order to use this variable for our analysis we transform it into a dummy variable yielding 4 dummies, one for each category.

Then, we focus our analysis on 12 variables to predict the price of a house. The control variables are then: *housing_median_age*, *total_rooms*, *total_bedrooms*, *population*, *households*, *median_income*, and the set of dummies *<1H.OCEAN* (1 for houses near the ocean by less than 1 hectare, 0 otherwise), *INLAND* (1 for properties in central areas, 0 otherwise), *ISLAND* (1 if a property is on an island, 0 o.w.), *NEAR.BAY*(1 if a house near the Bay of San Francisco, 0 otherwise), *NEAR.OCEAN* (1 if a House is Near the ocean, 0 otherwise). We consider *median_house_value* as the explanatory variable and note that this variable and *median_income* records the median values instead of means since usually for house prices and income there is a high positive skewness sometimes yielding to high outliers that may interfere with later statistics. So, for these variables is more appropriate to adopt a median measure to avoid the impact of these outliers.

We cleaned the data by dropping those missing values from the variable *total_bedrooms*. The missing data might be the case of districts such as an industrial district where the properties consists,

in majority, of warehouses. In those cases, we want to remove these observations since we are interested in analyzing residential properties.

To correct for skewness we apply the natural log to the variables *median house value*, *median income*. Also, since we suspect that houses tend to decrease its value with the past of time but at some points those old house may become expensive we apply a square root transformation to *median house age* in order to check if the assumption of decreasing at a increasing rate is applicable here. We note that some variables in the data set suffer of multicollinearity, *households* and *total_bedrooms* share a strong positivelinear relation (correlation of 0.98), this multicollinearity between these variables may produce biased coefficients when running OLS regressions. Three other cases are *households* and *total_rooms* (ρ = 0.92), *total_bedrooms* and *total_rooms* ($\rho = 0.93$), and *households* and *population* ($\rho = 0.91$). We do not want to omit these variables instead we use them to create rates per 100 individuals. We standarize these variables in order to compare districts with different population sizes. The introduction of *bedroom rate* $= \frac{total\ bedroom}{total\ rooms} * 100$ and *household rate* $= \frac{households}{total\ bedrooms} * 100$ reduces the problems of multicollinearity.

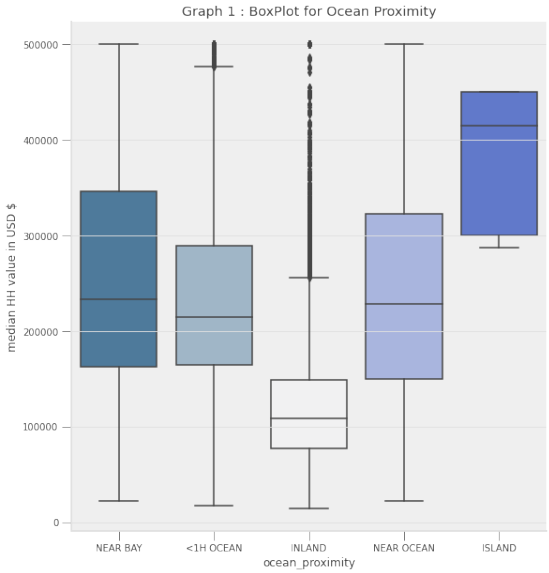
3 Summary Statistics.

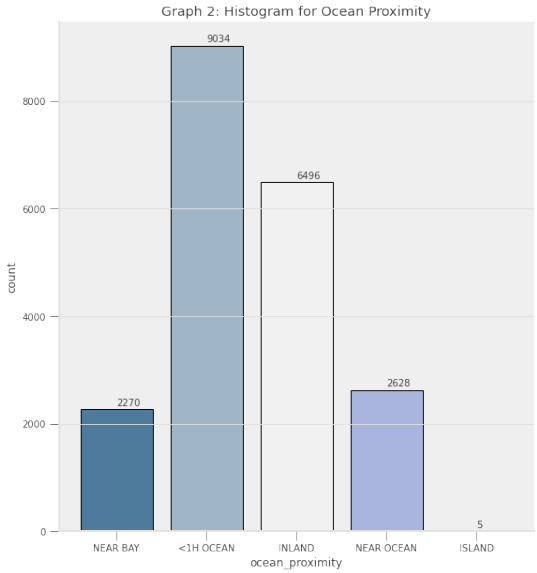
Table 1 shows the summary for the 20,433 observations in this sample, we can appreciate that the median house value oscillates between $\$14,999$ and $\$500,001$ with an average of $\$206,864.41$. This variable is widely dispersed, note that one standard deviation ($\$115,435.67$) is equivalent to half of the mean. The set dummies includes 9,034 houses located at less than one hectare of the ocean, the mean price of houses in this area is $\$240,268$. Inland records those houses that located more in the centre of California, it contains 6496 observations houses locate inland average a value of $\$124,896$. Near bay records 2270 houses near the bay of San Francisco, the average price in this area is $\$238,815$. Similarly, 2628 houses are located near the ocean, the average price here is $\$249,042$. The most expensive houses are located on islands, only 5 observations with average price of $\$380,440$. Graph 1 and graph 2 resume the findings on the dummies variables so far. The average median income reported for the households in this sample is \$3.87, but with a standard deviation of half the mean (\$1.89). For this variable the minimum value occurs at \$0.49) this means that the data is negatively skewed. Houses in the sample averaged an age of 28 years with

a standard deviation of 12 years, once again almost half the mean. There are new houses and houses as old as 52 years, very common in the real state. The average total rooms recorded 2636 rooms for an average number of households of 499 and and average population of 1424 people.

Table 1: Summary Statistics

	count	mean	std	min	max
median_house_value	20433.0	206864.413	115435.667	14999.0	500001.0
ln (house_value)	20433.0	12.085	0.569	9.616	13.122
median_income	20433.0	3.871	1.899	0.5	15.0
ln (income)	20433.0	1.245	0.471	-0.693	2.708
housing_median_age	20433.0	28.633	12.592	1.0	52.0
housing_median_age^2	20433.0	978.4	751.123	1.0	2704.0
total_rooms	20433.0	2636.504	2185.27	2.0	39320.0
total_bedrooms	20433.0	537.871	421.385	1.0	6445.0
bedroom_rate	20433.0	21.304	5.798	10.0	100.0
population	20433.0	1424.947	1133.208	3.0	35682.0
households	20433.0	499.433	382.299	1.0	6082.0
household_rate	20433.0	36.435	9.34	0.08	144.444





The given dataset is extensive and analyzing in order to find groupings of houses near the coast of California can be time consuming. Then, we rely on a visualization in order to have a quick glance, at the land level, of how the observations are distribution over a map. Since the dataset contains longitude and latitude for each observation we use them and plot the observations as dot on the map of California. The map 1 shows the location of the houses in the Map of California sorted by median house value. We sorted from the highest value to the lowest by quantiles, then we can appreciate that the houses in the 5th quantile are the expensive ones and are represented as yellow dots in the map 1. As suspected, the expensive houses are located near the ocean, however

we can see that this is not exclusive since there are some dots from the 4th quantile near the ocean and even a few from the 1st quantile. This is very interesting since we can deduce that even location near the ocean is a good predictor for the price of the house this is not the only factor that drives the price. While this graph is very informative we have two concentrations the first one near San Francisco and the second one near Los Angeles, so it may be helpful to see which location concentrates most of the houses, since it may be the case that the location with more houses turns to be the one with more population. Map 2 shows the concentration of the most expensive houses per km square in the given dataset. The are pretty packed near Los Angeles and with a lower concentration near San Francisco. Los Angeles and San Francisco host the most expensive houses in the dataset, however the concentration of these houses is more near the ocean that the bay in San Francisco so those houses near the ocean may not be only be expensive for being close the ocean but for being in a more concentrated location for houses. This finding is only possible by looking at the two maps, map 1 and map 2 together. Since, there are observations that are near the coast but are not expensive at all we are suspicious that others factors are driving the price down. We hypothesized that those houses are located near zones where earthquakes are common. We are going to use a website that provides information about earthquakes that took place in California from 1969 to 2007. The relevant years of information that we are going to use are in the range of 1969 to 1990. Note that we can use this information to sharpen our prediction of the the price of a House, that is, a house that is located over a geological failure or a well-known area of earthquakes may be cheaper that house located far away of this area. For the nature of the data the best approach is to use a map to check for this assumption. Map 4 shows the concentration of Earthquakes from 1969 to 1990 the interpretation of this map should not be confused with the magnitude of an Earthquake while there is a higher concentration of earthquakes in the north-east of the map there may be a single earthquake in any area with a higher magnitude. Overall, these two maps show that near the bay area there are a lot of earthquakes in comparison to Los Angeles area, this may be one of the reasons of a higher concentration of expensive properties near Los Angeles. However, is well know that Los Angeles is located over the Fault of San Andreas and the earthquakes follow a random pattern so there is no chance to predict where the next earthquake will take place. Finally, note that those residential buildings located near zones of high earthquakes activity may be expensive on average due to the materials used in their construction

(i.e. anti-seismic buildings). To analyze in further detail the magnitude of earthquakes we may need to add another map.

To sharpen our empirical analysis we add map 4, this map shows the earthquakes in California between 1969 to 1990 by its magnitude. Note that as previous suggested there are three areas of earthquakes concentration while the north- east area concentrates most of the seismic activity the most intense earthquake in this range of years took place near the San Francisco bay area. The north-west location is of our interest since it concentrates three intense earthquakes that my be derived in a tsunami affecting directly to the coast-population (maybe not the best place to purchase a house). Note that Los Angeles suffered two major earthquakes of around 6 - 6.5 in magnitude but in comparison with the rest of the state the earthquakes appear to not be common near this area.

4 Section Results

As previously stated, our aim is to test the hypothesis that those houses near the ocean tend to be more expensive on average than those that are not so close to the coast. Then, we run a number of regression in order to find the best model possible given this dataset. We define the median house price as the dependent variable and the set of dummies together with the set of control variables as the explanatory variables. Thus, we may note that the dummies act like averages for each category, similarly some X's are expected to hold a linear relationship with our dependent variable. That is the case of median income and median house price. Our first regression consists in a naïve model where we only use the median income of a person to predict the average price of a house. This simple regression yields biased results as one can appreciate in the high AIC and BIC values, similarly the {R^2} is quite low for a simple OLS model explaining only 47.38% of the variation in the price. Then, we run a second regression (model 2) where the controls with the dummies are added showing a positive bias in the median_income coefficient when these controls are omitted, the model improves slightly with a R squared adjusted of 63.43% and a slightly lower AIC and BIC values. We have to be suspicious for the reason that these controls do not increase notoriously our AIC and BIC values. If one simple looks at the coefficients p-values the we may conclude that the model is appropriate since all of the coefficients are significant at even less than

the 0.01% confidence level; however, the almost null variation in the AIC and BIC tends to show that there may be multicollinearity problems or there are outliers that affect our estimates. Then, we suggested taking the natural log of median_housing_price and median_income as a method to smooth the observations and reduce the impact of outliers.

Model 3 includes these transformations the AIC and BIC reduce substantially and the R-squared adj. increases this is an actual good sign for the regression model, we still think that we can made it perform better.

In table 2 we noted that many variables in this dataset share almost perfect linearity, this is a direct violation of the assumptions regarding the OLS estimation. Then, instead of using the variables total_bedrooms and households we suggest to take a rate of them per 100 individuals, in order to make the observations comparable even when they in reality may differ in magnitude. The transformations bedroom_rate and household_rate are then included into model 4 along with the variable population. We decide to include the variable population since, as table 2 suggest, there is no a high correlation between this variable and the transformations suggested. Finally, the model 4 is proposed as the best model possible given the data set, in equation form can be written as:

$$\widehat{Ln(median\_house\_value)} = \hat{B_0} + \hat{B_1}\, Ln(median\_income) + \hat{B_3}\, bedroom\_rate + \hat{B_4}\, household\_rate +$$

$$\hat{B_5}\, housing\_median\_age + \hat{B_6}\, housing\_median\_age^2 +$$

$$\hat{B_7}\, pop + \hat{B_8}\ < 1H.OCEAN + \hat{B_9}\, INLAND + \hat{B_10}\, ISLAND + \hat{B_{11}}\, NEAR\ BAY + \hat{B_{12}} NEAR\ OCEAN$$

The multivariate regression in model 4 has the lowest AIC and BIC with the highest Adj. R-squared among the four regressions here analyzed, then we choose this as the best model. This regression suggest that those properties located near the ocean have an average price of $USD\$34,544.37$. Similarly those properties less than one hectare near the ocean have an average price 1 percent lower than our base model (Near Ocean), properties located in land perform even worse with an average price 44 percent points lower and near bay only 4 percent lower that our base. We expect the averages prices of houses located on Island to be 60 percent higher than the average of the houses located near the ocean. Note that all of these estimates are considered

while controlling for the other variables. All the variable in dummies are significant at even less than the 1% level except for *<1H OCEAN* which is not significant at any significance level. This is pretty interesting, before including the transformation of model 4, model 3 contains all the coefficients significan at at least the 5% level, however the model 3 shares high levels of multicollinearity among the variables. Finally, when testing the hypothesis

$$\mathbf{H_0 : B_{dummy_i} = 0}$$

versus

$$\mathbf{H_1 : B_{dummy_i} \neq 0}$$

we reject the null hypothesis at even less than the 1% level and conclude that the location of the house in reference to the ocean plays a fundamental role in the price setting of a property for all the types of locations here presented except for "<1H OCEAN".

5 Conclusion

Through this paper we have used some statistical techniques in order to investigate any statistically significant relationship between the price of a House in 1990 in California and its location to the ocean. We use an extensive dataset from the California census in 1990 and a set of maps in order to identify any causal relationship. We found, by using maps, that there are observations that are located near the coast but are among the cheapest, we suggest that maybe it is due to their location to centres of intensive earthquake activity. The use of maps in order to analyze the real state market is pretty new and we decide to combine this information with web scrapping techniques in order to extract information about earthquakes. In fact we found, via maps, that houses located near those places that have been heavily struck by earthquakes tend to be cheaper. This type of empirical analysis done by maps is pretty new and must be assessed with statistical relevant information in order to find relevant findings, but still these types of techniques while basic is a good starting point. Thus, using the given dataset we found that those properties located near the ocean tend to be more expensive, on average, to those properties located far away of the coast even at the 1% level. Future research should consider to control for autocorrelation in the

error term expanding the basic econometrics models here presented to models that can control for this issue, sharpening the inference tests here presented. Adding more covariates is also recommended since can help to sharpen predictions. Collect data at the district level may be inefficient, as we can see there is difficult to distinguish exactly between the variables “NEAR OCEAN” and “<1 H OCEAN”. In order to avoid this in the future take observations at the county level since it allows to controls for unobservables at the country level. Note that we made a brief analysis of the price of a property and the earthquakes historical in California, future research should consider expand this topic since as previously mentioned those houses near these sites are cheap even if they are near the coast. Measure this type of risk may be difficult once again consider to use fixed effects dummies regression, or go even further with time effects, clustered standard errors at the county level.

6 References

Zietz, J., Norman Zietz E. and Sirmans, G.S. (2008). Determinants of House Prices: A Quantile Regression Approach. *Journal of Real Estate Finance and Economics*, 37: 317–333. DOI 10.1007/s11146-007-9053-7.

Harlfoxem. (2016, August 25). House Sales in King County, USA. Retrieved December 18, 2020, from https://www.kaggle.com/harlfoxem/housesalesprediction.

SOCR Data 021708 Earthquakes. (n.d.). Retrieved December 18, 2020, from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_021708_Earthquakes

7 Appendix

Table 2: Correlation Matrix for the Variables

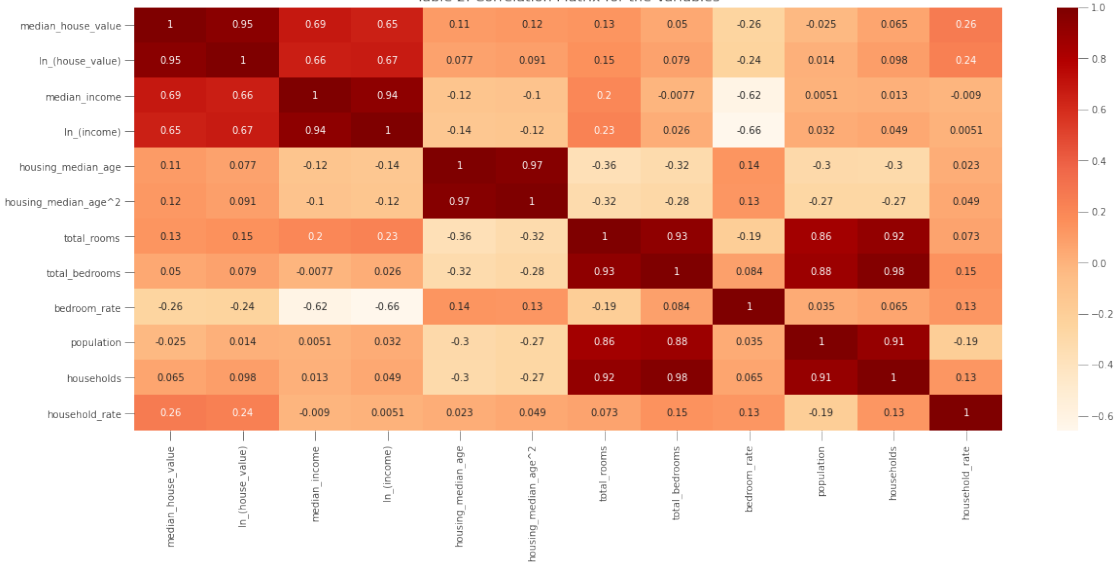


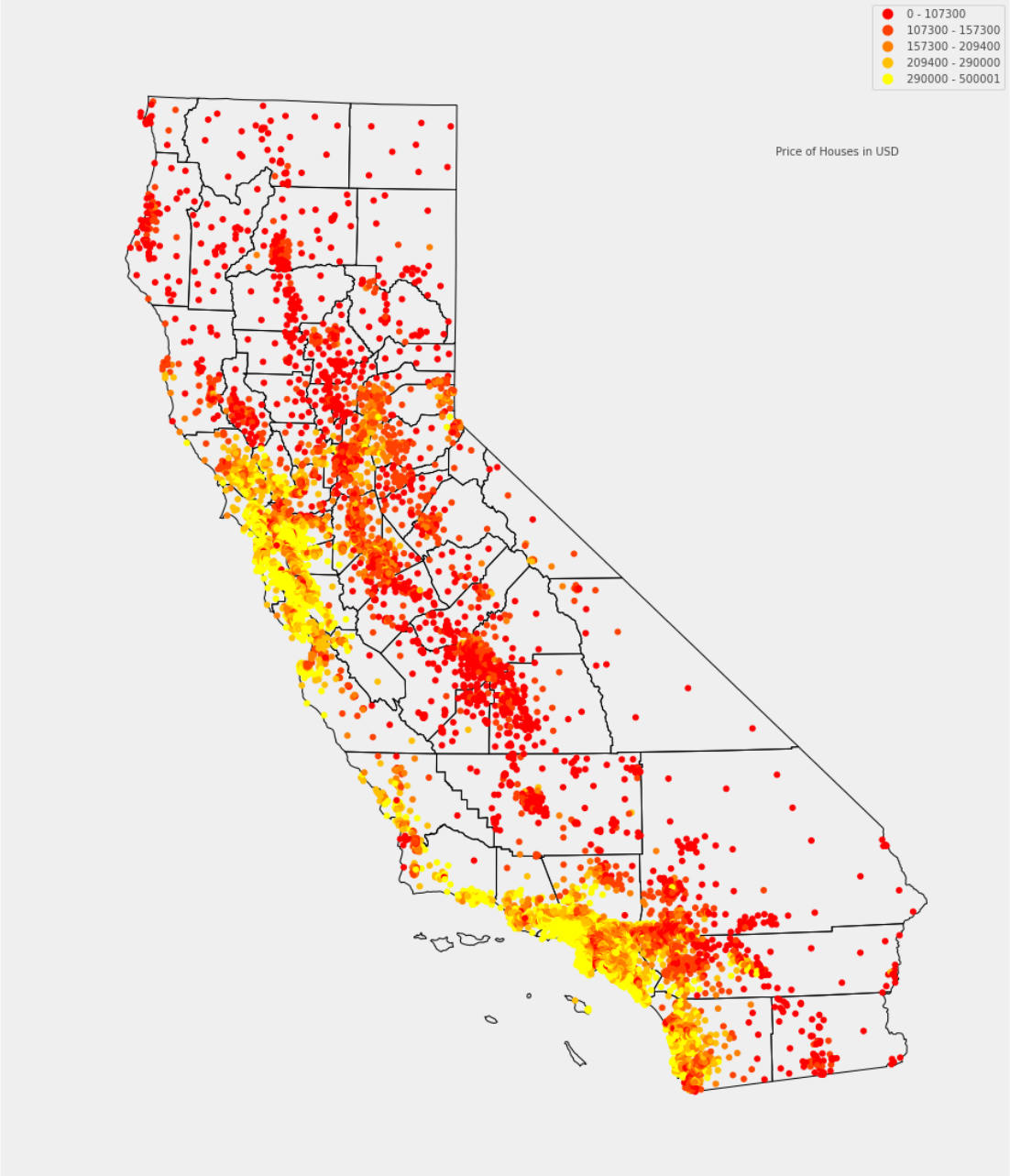
Table 3 - OLS Regressions

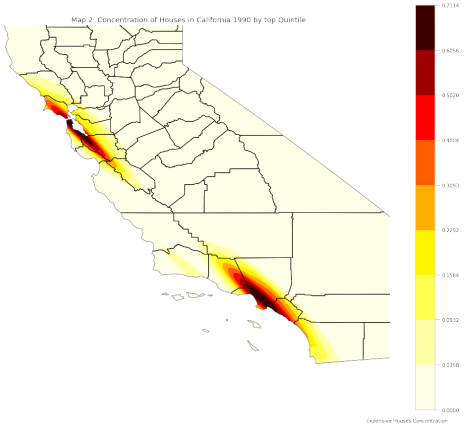
	Naive Model	Model 2	Model 3	Model 4
const	44906.37***	39474.04***	11.19***	10.45***
	(1329.97)	(2652.45)	(0.01)	(0.03)
<1H OCEAN		-13363.95***	-0.03***	-0.01
		(1556.18)	(0.01)	(0.01)
INLAND		-81389.10***	-0.50***	-0.44***
		(1675.77)	(0.01)	(0.01)
ISLAND		160788.39***	0.66***	0.60***
		(31263.10)	(0.15)	(0.14)
NEAR BAY		-8986.97***	-0.02**	-0.04***
		(2037.57)	(0.01)	(0.01)
R-squared	0.47	0.63	0.66	0.68
	0.47	0.63	0.66	0.68
No. observations	20433	20433	20433	20433
Adj R-squared	0.473808	0.634311	0.655762	0.682360
AIC	521221.802496	513795.562626	13179.753875	11536.649901
BIC	521237.652309	513882.736598	13266.927847	11623.823873

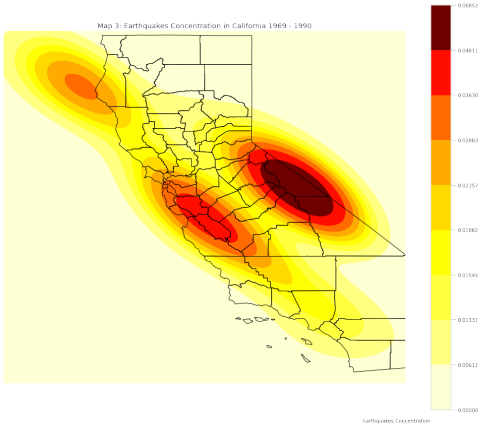
Standard errors in parentheses. The set of Covariates used are the same as the summary statistics table. Model 3 and 4 uses the Ln(median_house_value) and Ln(median_income) transformations. Model 4 uses bedroom rate and household rate instead of multicollinear variables.

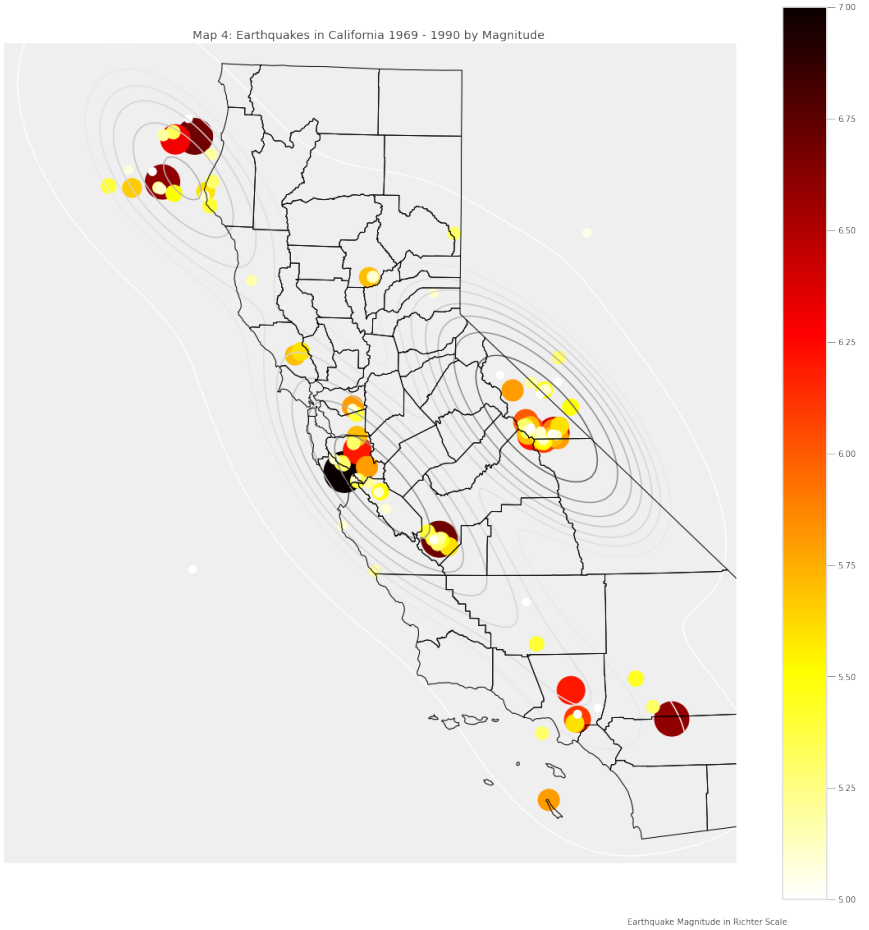
* p<.1, ** p<.05, ***p<.01

Map 1: California House Prices in 1990




Map 2: Concentration of Houses in California 1990 by top Quintile


Map 3: Earthquakes Concentration in California 1969 - 1990


Map 4: Earthquakes in California 1969 - 1990 by Magnitude
7.00
6.75
6.50
6.25
6.00
5.75
5.50
5.25
5.00
Earthquake Magnitude in Richter Scale